IN-65-CR

14423

P.49

Topics in Inference and Decision-Making with Partial Knowledge

S. Rasoul Safavian
David Landgrebe

TR-EE 90-53
September, 1990

School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

Table of Contents

Preface

Pattern recognition methods have become a common tool for analysis of Earth Observation multispectral image data. With the coming of the new, more complex sensors of the EOS system, it will be important to develop new enhancements to these tools in order that the full information-yielding capabilities of these new data be realized.

It is common in the theoretical derivation of a pattern recognition algorithm to assume precise knowledge of the parameters of the data. However, it is usually the case in the application of pattern recognition methods in practice that such precise knowledge is not available. For example, in order to obtain optimal performance from a Bayesian classifier, the *a priori* probabilities, the multivariate distributions, and appropriate loss functions for each class are needed; rarely is this information available in precise form. The question thus arises as to how best to model such imprecise knowledge and to modify the analysis scheme so that algorithms perform optimally under these more realistic circumstances. This question is what motivated this work.

Abstract[1]

Two essential elements needed in the process of inference and decision-making are prior probabilities and likelihood functions. When both of these components are known accurately and precisely, the Bayesian approach provides a consistent and coherent solution to the problems of inference and decision-making.

In many situations, however, either one or both of the above components may not be known, or at least may not be known precisely. This problem of partial knowledge about prior probabilities and likelihood functions is addressed. There are at least two ways to cope with this lack of precise knowledge: 1) robust methods, and 2) interval-valued methods.

First, ways of modeling imprecision and indeterminacies in prior probabilities and likelihood functions are examined; then how imprecision in the above components carries over to the posterior probabilities is examined. Finally, the problem of decision making with imprecise posterior probabilities and the consequences of such actions are addressed. Application areas where the above problems may occur are in statistical pattern recognition problems, for example, the problem of classification of high-dimensional multispectral remote sensing image data.

[1] Work reported here was supported in part by NSF Grant ECS 8507405 and NASA Grant NAGW-925

CHAPTER 1

1.1 Introduction

Inference is the process of observing a sample or samples and drawing information about certain parameters of the underlying process. There are two distinct categories to inference problems: some utilize prior information and others are based solely on the observation samples. It is taken as given that prior information should be used whenever available. To this extent the Bayesian approach provides a sound and coherent way of *combining* prior information, represented by prior probabilities and model information, represented by likelihood functions.

To put these matters in concrete terms, let us define $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$ as the set of parameters or the state of nature; $\pi(\theta_j)$ as the prior probability on Θ; and $\{p(x|\theta_i);\ \theta_i \in \Theta\}$ as the set of models or likelihood functions. Then after observing x, the inferential statement about θ_i is provided by the posterior probability $p(x|\theta_i)$ defined by the Bayes' formula

$$\pi(\theta_i|x) = \frac{p(x|\theta_i)\, \pi(\theta_i)}{\sum_{j=1}^{M} p(x|\theta_j)\, \pi(\theta_j)} \tag{1.1}$$

Decision-making problems are specific forms of inference problems. In decision making problems two other elements are added; namely a set of actions or decisions $\mathcal{D} = \{\delta_1, \ldots, \delta_n\}$ and a loss function $L(\theta_i, \delta_j(x))$. In many problems, the set of decisions and the set of parameters coincide. Then the problem of decision making is one of choosing an action from the set of actions or decisions $\mathcal{D}$, in such a way that the expected risk or the maximum risk is minimized.

1.2 Motivation for this research

As mentioned earlier, when all the components in the process of inference or decision making, namely the likelihood functions and the prior probabilities, are known the

Bayesian approach provides a consistent and coherent solution. In many real world problems, however, the above components may not really be known, or at least may not be known completely and precisely. For instance, in the early stages of outbreak of any new disease, with a small sample size, it is difficult if not impossible to obtain a precise model for the disease epistemology. Another example is the case of high sample dimensionality, where rarely if ever the available data is adequate to lead to a precise model.

The difficulty in specifying accurate prior probabilities is also very common. Actually the prior probabilities are often assigned quite subjectively. The difficulty in assigning accurate prior probabilities is the main reason non-Bayesian partisans attack the Bayesian approach. One can, however, go to the other extreme of doing away altogether with the prior probabilities. It seems self-evident that one should use all the information available without being either under- or over-committing.

1.3 Statement of the problem

The three interrelated problems to be addressed are: 1) how to describe imprecise prior probabilities and likelihood functions, 2) how to proceed from imprecise priors and likelihood functions to imprecise posterior probabilities, and 3) how to make decisions with imprecise posterior probabilities.

1.4 Useful concepts and terminologies

1.4.1 General remarks

It is important at this point to draw the differences between various sources of uncertainties; namely, randomness, vagueness, indeterminacies, etc. In this work, the main concern is with imprecisions resulting from one's inability to specify accurate priors and conditional densities. Therefore, imprecisions due to "indeterminacies" are the main concern.

An extreme case of indeterminacies is called "total ignorance". Conventional approaches for handling total ignorance (especially concerning prior probabilities) is to assign probabilities based on uniform distribution; i.e., if the state of nature is $\Theta = \{\theta_1, \theta_2,..., \theta_M\}$ and there is no prior information about the parameters, one may be inclined to assign,

$$\pi(\theta_i) = \frac{1}{M}, \quad i=1,\ldots,M. \tag{1.2}$$

There are at least two criticisms to this method of assigning probabilities:

1) In the case of "total ignorance", intuitively, probabilities should be assigned as

$$\pi(\theta_i) = [0,1], \quad i=1,\ldots,M.$$

2) When the state of nature Θ is continuous (e.g., $\Theta = \mathcal{R}$), this approach gives *improper* probabilities; i.e.,

$$\int_{\Theta} d\pi(\theta) = \infty \tag{1.3}$$

It is shown by Berger [3], that decisions based on improper distributions may give rise to inconsistencies (for definition, see below).

1.4.2 Terminology and Notation

The unknown quantity θ which affects the decision process is called the state of nature or the parameter. Prior probabilities for θ_i are denoted $\pi(\theta_i)$. The set of possible outcomes is the sample space and will be denoted $\mathcal{X}$. (Usually, $\mathcal{X}$ will be a subset of $\mathcal{R}^n$). The outcome of the experiment (i.e., the observation) will be denoted X. Often X will be a vector. The term "conditional densities", or "model", or "likelihood functions" is used to refer to the same quantity; i.e., $\{p(x|\theta_i); \theta_i \in \Theta\}$ or sometimes written as $\{p_{\theta_i}(x); \theta_i \in \Theta\}$. $E_\theta^x[f(x)]$ will denote the expectation (over X) of a function g(x), for a given value of θ. $L(\theta_i, \delta(x))$ will represent the losses incurred when upon observing sample x, decision $\delta(x)$ is made and the true state of nature is θ_i.

The risk of a decision rule $\delta(x)$ is defined as

$$R(\theta, \delta) = E_\theta^X[L(\theta, \delta(x)] = \int_{\mathcal{X}} L(\theta, \delta(x))\, dP(x|\theta)\,. \tag{1.4}$$

This is the expected loss, for each θ, if $\delta(x)$ is used repeatedly with varying x in the problem.

In order to decide about what type of decision rule should be used, some sort of *ordering* of decision rules is needed. The following definitions (Berger [3]) serve as guide lines.

DEFINITION 1.1: A decision rule δ_1 is *R-better* that a decision rule δ_2 if,

$$R(\theta, \delta_1) \leq R(\theta, \delta_2) \qquad \forall\, \theta \in \Theta. \tag{1.6}$$

DEFINITION 1.2: A decision rule is *admissible* if there exists no R-better decision rule. A decision rule is inadmissible if there exists an R-better decision rule.

DEFINITION 1.3: The *Bayes risk* of a decision rule δ, with respect to a prior distribution π on Θ, is defined as

$$r(\pi,\delta) = E^{\pi}[\, R(\theta, \delta)] = \int_{\Theta} R(\theta, \delta)\, \pi(\theta)\, d\theta\,. \tag{1.7}$$

Two frequently used decision-making principles are:

1) The *Bayes Risk* principle stated as

A decision rule $\delta_1$is preferred to a rule δ_2 if,

$$r(\pi, \delta_1) < r(\pi, \delta_2). \tag{1.8}$$

A decision rule that minimizes $r(\pi,\delta)$ is optimal and is called a Bayes rule.

2) The *minimax principle* stated as

A decision rule δ_1 is preferred to a decision rule δ_2 if,

$$\sup_{\theta} R(\theta, \delta_1) < \sup_{\theta} R(\theta, \delta_2)\,. \tag{1.9}$$

A decision rule is a minimax decision rule if it minimizes $\sup_{\theta} R(\theta,\delta)$ among all the rules in $\mathcal{D}$.

DEFINITION 1.4: coherency and coherent inference

The concept of coherency can be best explained in terms of betting, so let us define betting first.

> **DEFINITION 1.5** : A *bet* [46] concerning an event E is an arrangement whereby a sum of $\alpha\beta$ is exchanged for a sum of α if E occurs or 0 if it does not. The bet is said to be *on* or *against* E according $\alpha > 0$ or $\alpha < 0$. β is called the *betting rate* and α the *stake*. Let e be the indicator of the event E and 1 the indicator of the sure event Ω. Then a bet concerning E is a random quantity of the form $\alpha(e - \beta 1)$.

Let $(\Omega, \mathcal{A})$ be a measurable space with events $E_i \in \mathcal{A}$, $i = 1,2,\ldots,n$. And let a real-valued set function $P(E_i)$ represent the betting rate. Then De Finetti [7] shows that only when $P(E_i)$ is a probability, i.e. satisfies the axioms of probability, can one avoid the "sure loss" case. Only when P is a probability function, would it not be possible to select $E_1, E_2,\ldots, E_n$ and the stakes $\alpha_1, \alpha_2,\ldots, \alpha_n$ so that a combination of bets relative to these events, at the rates $P(E_i)$ for event E_i, i.e., $\sum_{i=1}^{n} (\alpha_i(e_i - P(E_i)\mathbf{1}))$, will assure a positive gain. (of course, it would be the same thing to require that no such quantity should be uniformly negative).

Decisions and inference based on coherent real-valued set functions, P, are called coherent decisions and inferences (Regazzini [33]).

CHAPTER 2

VARIOUS APPROACHES FOR HANDLING IMPRECISION

2.1 Introduction

The Bayesian approach offers an elegant way of combining prior information (i.e., prior probability π over Θ) and model information (i.e., $\{ p(x| \theta) ; \theta \in \Theta \}$), to construct a distribution p over $\Theta \times \mathcal{X}$; where p is the *unique* probability distribution over $\Theta \times \mathcal{X}$ that has π as its marginal for θ and the $p(x| \theta)$ as its conditionals given θ. After observing x, the Bayesian conditions p on x to obtain *posterior probabilities* for θ. Decisions based on the Bayes decision rule can be shown to be coherent.

The major criticisms to the Bayesian approach, however, are its requirements for *precise* knowledge of probability values and the *subjectiveness* of prior probabilities. The issue of prior probabilities being subjective is a philosophical one which will not be addressed here. One approach to make prior probabilities more *objective* (frequentist) is to obtain prior probabilities from n experts and use the weighted average of those.

In an attempt to relax the requirement for precise probability values, several methods have been proposed in the literature.

2.2 Minimum cross-entropy method

Many people [8,14,19,31,34,35,39-41,47] have tried to quantify available prior information and data without being *over committing*. One possible approach is the minimum cross-entropy method. Here, the prior information about an underlying distribution, p, and the available information I, which is usually in the form of constraints on the moments, is combined via operator $\circ$ to obtain the posterior probability q; that is [39]

$$q = p \circ I \tag{2.1}$$

Specifically, let q^* be the unknown underlying probability density function and the available information, I, be given as

$$\int g_k(x)\, q^*(x)\, dx = C_k \tag{2.2}$$

where $g_k(\cdot)$ are some known functions and C_k are known constants. Further, let the cross-entropy (also known as discrimination information, directed divergence, or I - divergence) between two probability density functions q and p be defined as

$$H[q,p] = \int q(x) \log[\, q(x)/p(x)\,]\, dx \tag{2.3}$$

Then a posterior probability $q(\cdot)$, whenever it exists, which minimizes the above quantity and satisfies the obvious restriction of

$$\int q(x)dx = 1 \tag{2.4}$$

is the one given by [39-41]

$$q(x) = p(x) \exp\left\{ -\lambda - \sum_k \beta_k\, g_k(x) \right\} \tag{2.5}$$

where β_k and λ are the Lagrangian multipliers for equations (2.2) and (2.4).

Remarks:

1) It has been shown [40], that the only operator o that satisfies uniqueness, invariance, and some other axioms of consistent inference and is implemented by means of functional analysis is the one given by the principle of minimum cross-entropy.
2) The maximum entropy method is a special case of the minimum cross-entropy method where there is no prior information or prior information is uniformly distributed.
3) Intuitively, the minimum cross-entropy method provides a posterior probability q(x) that is the closest distribution, in the sense of H[q,p], to the prior distribution p, yet satisfying the new information provided.
4) Even though H[q,p} is not a metric (does not satisfy the triangle inequality) it is a good information theoretic measure of closeness.
5) q is closer to the unknown underlying distribution q^* than is p.

The main difficulties with the minimum cross-entropy methods are [3]

1) In many cases a solution may not exist.

2) The requirement that information I be specified as various moments could be very restrictive.

3) A solution, when it exists, is usually in many senses non-robust.

2.3 "Sup" and "inf" approach

Let Ω be the sample space and $\mathcal{A}$ the appropriate σ-algebra on Ω. The most natural way to incorporate imprecision (i.e., indeterminacies) in probabilities is to define a family of probability measures $\mathcal{P}$, instead of a single probability measure p, over $(\Omega,\mathcal{A})$. This naturally leads to upper and lower probabilities

$$P^*(A) = \sup_{P\in\mathcal{P}} P(A) \qquad \forall A\in\mathcal{A} \tag{2.6}$$

and

$$P_*(A) = \inf_{P\in\mathcal{P}} P(A) \qquad \forall A\in\mathcal{A} \tag{2.7}$$

True probabilities, P(A), are upper and lower bounded as

$$P_*(A) \leqq P(A) \leqq P^*(A) \quad , \forall A\in\mathcal{A}. \tag{2.8}$$

Note that, even though every $P\in\mathcal{P}$ is a regular probability measure, P^* and P_* themselves need not be additive probabilities. Depending on the structure of $\mathcal{P}$, P^* and P_*may be measures that instead of being additive, are super- and sub-additive known as *Choquet capacities* ; capacities will be defined rigorously in the sequel.

2.4 Robust methods

The term "robust" was first used by Box in 1953. It usually refers to the situation where the performance does not degrade much as the parameters (here prior probabilities and likelihood functions) vary from their nominal values. There are two aspects to robustness; i.e. robustness analysis (also known as sensitivity analysis) and robustness design. The terms robust and *stable* are used sometimes to mean the same thing.

2.4.1 Distributionally robust approaches:

Here, first a set of nominal likelihood functions (or models) and a set of nominal prior probabilities is specified. One could do this even when the sample size is small and there is not much confidence in the sample. Then define a neighborhood for the nominal model and a neighborhood for the nominal priors. These neighborhoods reflect our confidence (or lack of it) in the nominal values. Finally, design the inference or decision-making procedure with the fact in mind that the actual model and the actual priors could vary within their respective neighborhoods. One could define these neighborhoods at least in two ways:

I) the neighborhood of a given model,

II) neighborhoods composed of a mixture of models.

I) The neighborhood of a given model M_0 :

Let M be the class of all models (e.g., the class of all prior probabilities, or the class of all likelihood functions). Let M_0 be the nominal model and M_1 be a wider class of models including $M_{0.}$ This idea is easily depicted in the following figure

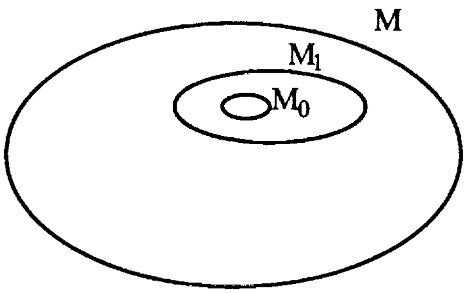

Fig. 1

II) The neighborhood composed of a mixture of models:

When it is difficult to justify a single neighborhood of a model, one defines neighborhoods that are composed of a mixture of models. Graphically this is illustrated in the following fig. 2

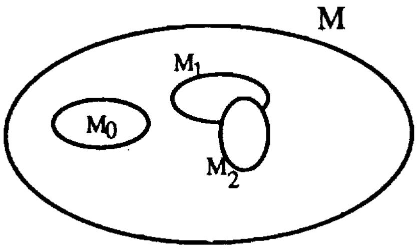

Fig. 2

Specific examples of these neighborhoods for prior probabilities and likelihood functions follow.

2.4.2 Example of neighborhoods for prior probability

Let us assume that π, the true prior probability, belongs to class Γ of the prior probabilities, where Γ is defined as follows:

1) **Band model:** [11]

$$\Gamma = \{ \alpha\pi : \ L \leqq \pi \leqq U \} \qquad (2.9)$$

where L and U are lower and upper nonnegative bounding functions and α is just a normalizing factor to make π a probability measure.

One way to obtain lower and upper bounds is to estimate the prior probabilities and then find a confidence interval (limit) for the estimates; thus creating a band for priors.

Remark:

Strictly speaking, prior probabilities should be independent of data and should be provided without looking at the data.

2) **ε - contamination model:** [34]

$$\Gamma = \{ \pi : \pi = (1\text{-}\varepsilon)\, \pi_o + \varepsilon\, \pi_1 \} \qquad (2.10)$$

where π_0 is the nominal prior probability, ε is the degree of uncertainty in the nominal priors, $0 \leq \varepsilon \leq 1$, and π_1 is any unknown and completely arbitrary probability measure.

The rationale for this model [18] is as follows. Consider a Bayesian decision-maker who after looking at the observation x realizes that his prior belief π was very far off the mark. Should he stick to it and obtain a posterior distribution nobody, even he himself will believe in? Or should he cheat and change the prior? ε-contamination allows one to keep an ε of the prior mass in "reserve for emergencies" to cope with situations like above.

3) **Prior probabilities specified by linear inequalities:** [31]

In many cases one may only be able to make statements such as: θ_1 is ten times more likely than θ_2, or θ_1 is less likely to occur than θ_2 and θ_3, etc. Such partial prior information could be specified by set of linear inequalities of the form

$$\Gamma = \{ \pi : \alpha\pi \geqq 0 , \mathbf{1}^T\pi = \mathbf{1} , \pi \geqq \mathbf{0} \} \tag{2.11}$$

2.4.3 Uncertainty models for likelihood functions

In many cases, a precise model for the phenomena under observation may not be available. For instance, in the *early stages* of a new disease a precise model may not be available. Obviously, it would not be very appropriate to use a very precise model since consequences of error in the assumed model may be very serious both financially and in terms of human factor. Two extreme case approaches here are either adapting parametric approaches or distribution-free approaches. Something between these two extreme cases will be raised, however.

1) **Elaborated model:** [15]

Let $f(x|\theta)$ be the nominal model for data x and parameter θ. Then an elaborated model (EM) can be represented as a family of densities $\{f(x|\theta, \lambda), \lambda \in \Lambda\}$ with $f(x|\theta) = f(x|\theta, \lambda_o)$ for some $\lambda_o \in \Lambda$. Examples of this type of model are:

1.1) The exponential power family

$$f(x| \mu,\sigma,\lambda) \propto \sigma^{-1} \exp\left\{ -c(\lambda) \left| \frac{x-\mu}{\sigma} \right|^{\left\{\frac{2}{1+\lambda}\right\}} \right\} \tag{2.12}$$

with $\lambda \in (-1,+1)$. Here $\lambda \to -1$ corresponds to a uniform density, $\lambda = 0$ to the normal density, and $\lambda = +1$ to the double exponential density. λ could be considered, here, as a measure of kurtosis.

1.2) The Huber family

$$f(x|\mu,\sigma,\lambda) \propto \exp\left\{ -g\left(\frac{x-\mu}{\sigma} \right)\right\} \tag{2.13}$$

where

$$g(x) = \begin{cases} \frac{1}{2}x^2 & |x| < \lambda \\ \lambda|x| - \frac{1}{2}\lambda^2 & |x| > \lambda \end{cases} \tag{2.14}$$

Notice that as $\lambda \to 0$, this becomes a double exponential density and as $\lambda \to \infty$, it tends to the normal density. For other values of λ, one obtains a normal center and exponential tails. Notice that to proceed with the above model to the posterior probabilities one would require the knowledge of the joint prior probabilities, $\pi(\theta,\lambda)$. This point will be returned to latter.

2) **Band model:** [20]

Conceptually, band models for likelihood functions (or conditional densities) are similar to the band models for prior probabilities. Suppose $f_\theta(x)$ (or $f_\theta(x|\theta)$) is the density function, with respect to some measure μ (e.g., Lebesgue measure) on the measurable space $(\mathcal{X},\mathcal{A})$, of a probability measure $P_\theta(x)$ (or $P_\theta(x|\theta)$) Consider the neighborhood defined as,

$$\mathcal{F} = \left\{ f \mid f_L \leq f \leq f_U ,\ \int_{\mathcal{X}} f \, d\mu = 1 \right\} \tag{2.15}$$

where f_L,f_U are nonnegative bounding functions with f_L being bounded. This model may be useful, for instance, when the density function f estimated from training data are expressed as lying within pairs of confidence limits.

3) **ε – contamination model**: [22]

$$\mathcal{F}_i = \{ f_i \mid f_i = (1-\varepsilon_i)\, f_o + \varepsilon_i\, h_i ,\ h_i \in \mathcal{H}_i \}. \tag{2.16}$$

This is also similar to the model introduced above for prior probabilities, except here f, f_o, and $\mathcal{h}$ are conditional densities. This model was first introduced by Tukey and has the following intuitive interpretation in statistical classification problems.

The class θ_i of observations consists of two classes: the well known frequently observable class that has the known density f_o and the non-studied, rarely observable class with an unknown density h_i; $h_i \in \mathcal{H}_i$. If θ_i is observed, then an observation from the first part appears with probability (1- ε_i) and from the second part with probability ε_i. The ε-contamination model is a special case of the band model where $f_{i_L} = (1-\varepsilon_i)f_i$ and $f_{i_U} \rightarrow \infty$.

4) **Total variational model**: [34]

This is another useful neighborhood defined as

$$\mathcal{P} = \{ P : | P(A) - P_o(A) | \leq \varepsilon \} , \quad \forall A \in \mathcal{A} \tag{2.17}$$

where $(\Omega, \mathcal{A})$ is the measurable space on which the probability measures are defined. In terms of densities, it can be written as

$$\mathcal{F} = \left\{ f : \int | f(x) - f_o(x) | \, dx \leq \varepsilon \right\} \tag{2.18}$$

Once the neighborhoods are defined, then the problem of decision making is to choose an action, from the set of possible actions (or decisions) that minimizes the maximum risk; i.e., a minimax approach. Let us use the notation introduced earlier; except to make things more explicit, the risk function will be written as

$$r(\pi, \delta) = r(\pi(\theta), \{f_\theta(x)\}, \delta(x)) \tag{2.19}$$

Then the minimax decision rule δ^*, (or actually, Γ- minimax decision-rule, since the priors are allowed to vary too) is given by

$$\delta^*(x) = \arg \min_{\delta \in \mathcal{D}} \left(\max_{\substack{\pi \in \Gamma \\ \{f_\theta\} \in \mathcal{F}}} r(\pi(\theta), \{f_\theta\}, \delta(x)) \right) \tag{2.20}$$

The prior probability π and $\{f_\theta\}_{\theta\in\Theta}$ for which δ^* is attained are called the *least favorable* distributions and would be denoted $(\pi^L(\theta), \{f_\theta\}^L)$. Note that δ^* and $(\pi^L(\theta), \{f_\theta\}^L)$ satisfy

$$r(\pi(\theta), \{f_\theta\}, \delta^*(x)) \leqq r(\pi^L(\theta), \{f_\theta\}^L, \delta^*(x)) \leqq r(\pi^L(\theta), \{f_\theta\}^L, \delta(x)) \quad (2.21)$$

It is important to note here that, even though it is conceptually simple to model above minimax approach, obtaining solutions (i.e., $\pi^L(\theta)$, $\{f_\theta\}^L$,δ^*) may not be so simple. Solutions for minimax (but not Γ-minimax) problems have already been found for certain type of neighborhoods such as ε-contamination, band models and total variational neighborhoods. These neighborhoods all have one thing in common: They all could be specified as $\mathcal{P}$, where $\mathcal{P}$ is a family of distributions defined over measurable space $(\Omega,\mathcal{A})$ as

$$\mathcal{P} = \{ P\in \mathcal{M} : \ P(A) \leqq v(A) \ , \ \forall A\in \mathcal{A} \} \quad (2.22)$$

where $\mathcal{M}$ is the set of *all* probability measures defined over measurable space $(\Omega,\mathcal{A})$. $\mathcal{P}$ is said to be the set of probabilities *majorized* by v.

For an *ε-contamination* neighborhood, $v(A)$ is defined as

$$v(A) = (1-\varepsilon)\, P_o(A) + \varepsilon \qquad , A\neq \emptyset \quad (2.23)$$

For a *total-variational* neighborhood, $v(A)$ is defined as

$$v(A) = \min\left(P_o(A) + \varepsilon , 1 \right) \qquad , A\neq\emptyset \quad (2.24)$$

$v(A)$s defined above have an interesting property; namely, they are set functions that satisfy the following properties [5]

p1) $v(\emptyset) = 0, \qquad v(\Omega) = 1$

p2) $A\subset B \quad \Rightarrow \quad v(A) \leqq v(B)$

p3) $A_n \uparrow A \quad \Rightarrow \quad v(A_n) \uparrow v(A)$

p4) $F_n \downarrow F, F_n$ closed, $\Rightarrow v(F_n) \downarrow v(F)$

and

p5) $v(A\cup B) \leqq v(A) + v(B) - v(A\cap B)$.

Any set function that satisfies p1)-p4) is called a Choquet capacity or *capacity* for short. If it also satisfies p5), then it is called alternating of order 2, or for short, *2-alternating* capacity. A set function u that satisfies p1)-p4), and instead of p5) satisfies

$$\text{p6)} \quad u(A \cup B) \geqq u(A) + u(B) \cdot u(A \cap B)$$

is called a *2-monotone* capacity. More generally, consider the successive differences defined [29] as

$$\nabla_1(B;B_1)_v = v(B) - v(B \cup B_1) \tag{2.25}$$

$$\nabla_{n+1}(B; B_1,\ldots, B_{n+1})_v = \nabla_n(B; B_1,\ldots, B_n)_v - \nabla_n(B \cup B_{n+1}; B_1,\ldots, B_n)_v \tag{2.26}$$

If $\nabla_k \leq 0$ for $k=1,\ldots,n$, then v is called an n-alternating capacity; if $\nabla_n \leq 0$ for all n, it is called and infinite alternating capacity. Similarly, let

$$\Delta_1(B; B_1)_u = u(B) - u(B \cap B_1) \tag{2.27}$$

$$\Delta_{n+1}(B; B_1,\ldots, B_{n+1})_u = \Delta_n(B; B_1,\ldots, B_n)_u - \Delta_n(B \cap B_{n+1}; B_1,\ldots, B_n)_u \cdot \tag{2.28}$$

If $\Delta_k \geq 0$ for $k=1,\ldots,n$, then u is called an n-monotone capacity; if $\Delta_n \geq 0$ for all n, u is said to be an infinite monotone capacity. Note that alternating and monotone capacities v and u, satisfy

$$v(A) + u(A^c) = 1 \tag{2.29}$$

and are said to be *conjugates*.

Let us consider the simplest form of decision-making; that is testing a null hypothesis H_0 versus an alternative hypothesis H_1. And suppose the prior probability of H_0 (and H_1) is known and is given by $\frac{t}{t+1}$ (and $\frac{1}{t+1}$), $t \in [0,\infty]$. Furthermore, suppose the hypotheses correspond to two imprecisely known likelihood functions; and they can be modeled as sets majorized by 2-alternating capacities v_0 and v_1. That is,

$$\mathcal{P}_o = \{ p : p \leqq v_o \} \tag{2.30}$$

and

$$\mathcal{P}_1 = \{ p : p \leqq v_1 \} \tag{2.31}$$

Recall that this includes such models as the *ε-contamination* and the total-variational model, etc. Thus one is testing *composite* hypotheses

$$\begin{array}{l} H_o : X \sim \mathcal{P}_o \\ \text{vs.} \\ H_1 : X \sim \mathcal{P}_1 \end{array} \tag{2.32}$$

Let A be the critical region of test; i.e. reject $\mathcal{P}_0$ if $x \in A$ is observed. Then the *upper Bayes risk* of the critical region A is (Huber & Strass[17])

$$G_t(A) = \frac{t}{t+1} v_0(A) + \frac{t}{t+1} \left(1 - u_1(A)\right) \tag{2.33}$$

To minimize $G_t(A)$, it is enough to minimize the 2-alternating function

$$W_t(A) = t\, v_0(A) - u_1(A) . \tag{2.34}$$

Huber and Starssen [17] state and prove the following lemma

Lemma 1: For each $t \in [0,\infty]$ (i.e., any given priors), there is an A_t such that,

$$W_t(A_t) = \inf_{A} W_t(A) \tag{2.35}$$

Note that A_t minimizes the maximum Bayes risk.

Another approach, other than this minimax approach, could be one based on translating the imprecision in priors and likelihood functions onto posterior probabilities obtaining a family of lower and upper posteriors. For the sake of simplicity, let us examine the cases of imprecision in priors and likelihood functions separately.

First, let us assume that likelihood functions, p(x|θ), are known precisely and the only source of imprecision is due to priors which can also be modeled by the ε-contamination neighborhood

$$\Gamma = \{ \pi : \pi = (1\text{-}\varepsilon)\, \pi_o + \varepsilon\, \pi_1 \} \tag{2.36}$$

Let $\pi(\theta_i|x)$ denote posterior probability given by the Bayes rule as

$$\pi(\theta_i|\, x) = \frac{p(x|\theta_i)\; \pi(\theta_i)}{\sum_{\theta_i \in \Theta} p(x|\theta_i)\; \pi(\theta_i)} \tag{2.37}$$

and let $\pi_o(\theta_i|x)$ denote the posterior probability corresponding to the nominal prior π_o. Then Huber [18] shows that

$$\sup_{\pi \in \Gamma} \;\pi(\theta_i|\, x) \;= \frac{\pi_o(\theta_i|\, x) + S(\theta_i)}{1 + S(\theta_i)} \tag{2.38}$$

and

$$\inf_{\pi \in \Gamma} \;\pi(\theta_i|\, x) = \frac{\pi_o(\theta_i|\, x)}{1 + S(\theta_i^c)} \tag{2.39}$$

where

$$S(\theta_i) = \frac{1}{1\text{-}\varepsilon} \; \frac{p(x|\theta_i)}{\sum_{\theta_i \in \Theta} p(x|\theta_i)\; \pi_o(\theta_i)} \quad . \tag{2.40}$$

Now suppose that both the likelihood functions and the prior probabilities are given by the ε-contamination models. Following Huber [18], one says upon observing x, the "information" about Θ is increased by the (possibly negative) amount

$$\sum_{\theta_i \in \Theta} \pi(\theta_i|\, x) \log \pi(\theta_i|\, x) \; - \sum_{\theta_i \in \Theta} \pi(\theta_i) \log \pi(\theta_i) \quad . \tag{2.41}$$

Then a family $\{p(\cdot\, |\theta_i)\}$ of conditional densities and a prior probability π will be least informative if they minimizes

$$H(p, \pi) = E_x \,[\; \{ \sum_{\theta_i} \pi(\theta_i \,|\, x) \log \pi(\theta_i \,|\, x) \} \; - \sum_{\theta_i} \pi(\theta_i) \log \pi(\theta_i) \;] \tag{2.42}$$

$$= \sum_{x} \sum_{\theta_i} p(x \mid \theta_i)\, \pi(\theta_i) \left[\log \frac{p(x|\theta_i)\, \pi(\theta_i)}{\sum_{\theta'_i} p(x|\theta_i)\, \pi(\theta'_i)} \right] - \sum_{\theta_i} \pi(\theta_i) \log \pi(\theta_i) \quad (2.43)$$

subject to the side conditions that

$$\sum_{x} p(x| \theta_i) = 1 \ . \quad (2.44)$$

and

$$\sum_{\theta_i} \pi(\theta_i) = 1. \quad (2.45)$$

Note that it was assumed that $\mathcal{X}$ is finite. Solution for this problem, except perhaps for very trivial cases, is difficult to obtain.

2.5 Interval-valued probabilities

Bayesian frame of inference and decision-making requires precise probabilities and has no provisions for imprecise knowledge. There has been many attempts [2,8,11,24,42,43, 45,46,47] to generalize classical "point-valued" probabilities to "interval-valued" probabilities. Dempster [8-10], and later Shafer [35-38], in an attempt to generalize the Bayesian framework, have come up with what is known as the Dempster-Shafer (D-S) theory of evidence [35]. We will start with an example first, then proceed to point out the major problems with the D-S theory, and finally describe a more natural extension of usual probability measures and Bayes theorem cast in this new framework.

2.5.1 Dempster-Shafer theory

The basic idea can become clear with the following (desk) example. Suppose there is a desk with two drawers on the right side: the right top drawer (RT) and the right bottom drawer (RB). There are three drawers on the left side: the left top drawer (LT), the left middle drawer (LM), and the left bottom drawer (LB). Suppose a file is placed, at random, in one of the drawers. Further suppose that the available information (evidence in the D-S language) is given as

$$\text{Prob (file is in the left side drawers)} = m_1 = 0.5 \qquad (2.47)$$
$$\text{prob (" " " " (RT) ")} = m_2 = 0.2$$

and there is no more information.

Note that the total evidence, $m_1 + m_2 = 0.7 < 1$. Shafer calls the difference (1– 0.7 =0.3), the global ignorance. The global ignorance can be assigned to any of the drawers, and yet none in particular. Then given the above scenario, one would like to answer questions like what is the probability that the file is in the (LM) drawer? etc.Obviously, the answer to this question can not be given by a single number. George Boole [4] was the first to realize this point and he suggested the idea of inner and outer measures, p_* and p^*, such that probability of any event, p, is bounded by p_* and p^* as

$$p_* \leq p \leq p^* . \qquad (2.49)$$

Then how does one compute p_* and p^*? Shafer calls m's the *basic probability assignments* or (bpa)'s. m(A) represents the measure of belief that is committed *exactly* to set A and not to any of its proper subsets. Moreover, let us denote the sample space by Ω, and assume it is finite. Let 2^Ω represent the power set of Ω. Then

DEFINITION 2.1: (Shafer [35])
A function m: $2^\Omega \rightarrow [0,1]$ is called a basic probability assignment (bpa) whenever

$$(1) \quad m(\emptyset) = 0 \qquad (2.50)$$

and

$$(2) \quad \sum_{A \subset \Omega} m(A) = 1 . \qquad (2.51)$$

Note that

i) It is not required that $m(\Omega) = 1$;

ii) It is not required that $m(A) \leq m(B)$ when $A \subseteq B$;

iii) There is no obvious relationship between m(A) and $m(A^c)$.

Recall that m(A) reflects the measure of belief that is committed exactly to A, not the *total* belief that is committed to A. To obtain the total belief committed to A, Shafer argues, that one must add to m(A), the bpa of all the proper subsets B of A. He calls this "BELIEF" or Bel for short. That is

$$\mathrm{Bel}(A) = \sum_{B \subseteq A} m(B) . \qquad (2.52)$$

Dempster in his original work called these Bel's, lower probabilities. More formally, a function Bel: $2^{\Omega} \rightarrow [0,1]$ is called a belief function over Ω if it is given by (2.52), for some bpa m: $2^{\Omega} \rightarrow [0,1]$. For our earlier "desk" example :

Bel (file is in (ML) drawer) = 0.

Bel(file is in (RT) drawer) = 0.2.

It is important to note that

$$\mathrm{Bel}\,(A) + \mathrm{Bel}\,(A^c) \leq 1 . \qquad (2.53)$$

To see the implication of this relationship, suppose there is no evidence at all to support A, or Bel(A) = 0. Then, (2.53) says that, in D-S theory, it is not automatically implied that $\mathrm{Bel}(A^c) = 1$; i.e., lack of belief in something does not necessitate its compliment.

Furthermore, the bpa that produces a given belief function can be *uniquely* recovered from the belief function. This inverse relation is called *mobius* inverse. For any belief function Bel, a dual function plausibility (or "Pl" for short) is defined as

$$\mathrm{Pl}\,(A) = 1 - \mathrm{Bel}\,(A^c) . \qquad (2.54)$$

In terms of bpa, m, plausibility could be written as

$$\mathrm{Pl}\,(A) = \sum_{B \cap A \neq \emptyset} m(B) . \qquad (2.55)$$

Dempster called these Pl's, upper probabilities. Note

$$\mathrm{Pl}\,(A) + \mathrm{Pl}\,(A^c) \geq 1 \qquad (2.56)$$

and

$$\mathrm{Pl}\,(A) \geq \mathrm{Bel}\,(A) . \qquad (2.57)$$

From our earlier "desk" example:

Pl (file is in (ML) drawer) = 0.3

Pl (file is in (RT) drawer) = 0.5.

To make the idea of "Bel" and "Pl" clearer, let us consider the following example. Suppose we are given: $m(B_1) = 0.3$, $m(B_2) = 0.4$, $m(B_3) = 0.1$, $m(\Omega) = 0.2$, and want to find the lower and upper probability (or Bel and Pl) of a set A given in the following diagram.

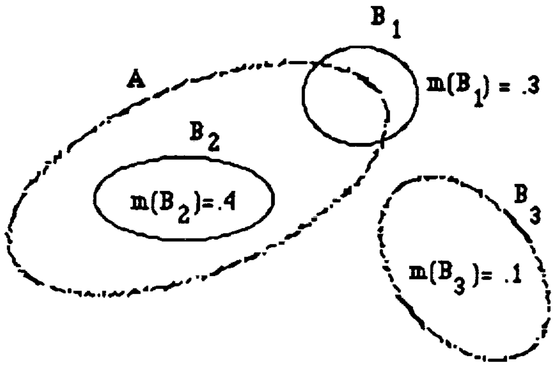

Fig. 3

Then

$$\text{Bel}(A) = \sum_{B_i \subseteq A} m(B_i) = m(B_2) = .4$$

$$\text{Pl}(A) = \sum_{B_i \cap A \neq \emptyset} m(B_i) = m(B_1) + m(B_2) + m(\Omega)$$

$$= .3 + .4 + .2 = .9 .$$

Shafer, further argues that the class of belief functions can be characterized without reference to basic probability assignments. That is:

THEOREM 2.1: Shafer [35]

A function Bel: $2^\Omega \rightarrow [0,1]$ is a belief function if and only if it satisfies the following:

(1) Bel $(\emptyset) = 0$.

(2) Bel $(\Omega) = 1$.

(3) for every positive integer n and every collection $A_1, A_2, ..., A_n$ of subsets of Ω

$$\text{Bel}(A_1 \cup .... \cup A_n) \geq \sum_i \text{Bel}(A_i) - \sum_{i<j} \text{Bel}(A_i \cap A_j) + ...+ (-1)^{n+1}\text{Bel}(A_1 \cap ..., \cap A_n).$$

Remark: Note that Bel functions are infinite monotone capacities.

Similarly, one can define plausibility functions as

THEOREM 2.2:

A function Pl: $2^{\Omega} \rightarrow [0,1]$ is a plausibility function if and only if it satisfies the following conditions:

(1) Pl $(\emptyset) = 0$.

(2) Pl $(\Omega) = 1$

(3) For every positive integer n and every collection $A_1, \ldots, A_n$ of subsets of Ω

$$\mathrm{Pl}\,(A_1 \cap \ldots \cap A_n) \leq \sum_{i} \mathrm{Pl}\,(A_i) - \sum_{i<j} \mathrm{Pl}\,(A_i \cup A_j) + \ldots + (-1)^{n+1}\,\mathrm{Pl}\,(A_1 \cup \ldots \cup An\,).$$

Remark :

1) Note that Pl functions are infinite alternating capacities.

2) When $\mathrm{Bel}(A \cup B) = \mathrm{Bel}(A) + \mathrm{Bel}(B)$, $A \cap B = \emptyset$ belief function becomes the usual classical probability measures. Furthermore, one can show that (Klir [23]) a belief function, Bel, on a finite power set 2^{Ω} is a probability measure if and only if its basic probability assignment, m, is given by $m(\{W\}) = \mathrm{Bel}(\{w\})$ and $m(\{A\}) = 0$ for all subsets of Ω that are not singletons.

3) A Bel function that satisfies Bel $(A) = 0$ for every proper subset A of Ω is called a *vacuous* belief function. In terms of basic probability assignments, this means $m(\Omega) = 1$ and $m(A) = 0$ for every proper subset A of Ω. Furthermore plausibility of every such A is one. That is

$$\mathrm{Bel}\,(A) = 0 \;\leq\; \mathrm{pr}\,(A) \;\leq\; \mathrm{Pl}\,(A) = 1 \qquad \forall\, A \subset \Omega\,.$$

Now that we are equipped with the basic notions of D-S theory, let us see how this theory address two major issues: 1) combination of various sources of information (evidence), and 2) the rule of conditioning.

2.5.2 Combination of various sources of information

First of all D-S theory requires that sources of evidence be *independent* (or non-interacting). Sources of evidence in remote sensing could be for instance, multispectral data, elevation data, slope data,etc. Or in medical diagnosis, sources could be the opinion of several doctors (experts) about the same patient. D-S theory proceeds to attain the bpa's from each source and then combines the bpa's with what is known as Dempster's *orthogonal sum*. More specifically, let Ω be the sample space and let m_1 and m_2 be two bpa's obtained from information sources S_1 and S_2 respectively. Then the total information obtained about Ω from the sources S_1 and S_2 is given by the new bpa m(c), given as

$$m(c) = \left(m_1 \oplus m_2\right)(c) = \frac{\sum_{A_i \cap B_j = c} m_1(A_i).m_2(B_j)}{1 - \sum_{A_i \cap B_j = \varnothing} m_1(A_i)\, m_2(B_j)} \tag{2.58}$$

Note that order of combination is not important. That is

$$m_1 \oplus m_2 = m_2 \oplus m_1 \tag{2.59}$$

i.e., Dempster's orthogonal sum is commutative. Also, if there are three independent sources specified by their bpa'sm_1, m_2, and m_3, they can be combined by the successive application of above rule. That is

$$m = \left(m_1 \oplus m_2\right) \oplus m_3 = m_1 \oplus \left(m_2 \oplus m_3\right) \tag{2.60}$$

and the order of aggregation is not important; i.e., $\oplus$ is an associative operator.

Intuitively, Dempster's orthogonal sum says that, to find the joint bpa for a set c, take all the sets from source S_1, i.e., A_i's, and all the sets from source S_2, i.e., B_j's, multiply their bpa's and add over all such sets. The denominator is a *normalizing* constant; it is required since one of the requirements for a valid bpa function is that it must sum to one. Dempster's orthogonal sum is the heart of D-S theory and also the major source of controversy and criticism. The following example, originally due to Zadeh [48-50], highlights this issue. Suppose $\Theta=\{\theta_1,\theta_2, \theta_3\}$ is the sample space of outcomes, and the information available from two independent sources lead to two sets of bpa's given below

	θ_1	θ_2	θ_3
m_1	.9	0.1	0
m_2	0	0.1	.9

Then upon applying Dempster's rule of combination, one obtains,

	θ_1	θ_2	θ_3
$m = m_1 \oplus m_2$	0	1	0

That is, even though both sources *individually* reflect low beliefs on $\{\theta_2\}$, after combination, they collectively confirm $\{\theta_2\}$! This is highly counter-intuitive; and again results from the normalization needed in the Dempster's rule.

Walley [44], Krantz and Miyamoto [25], and Shafer [36] have tried to apply the D-S theory to the problem of statistical inference. For the sake of simplicity, suppose that the state of nature Θ is finite; i.e., $\Theta = \{\theta_1, \ldots, \theta_n\}$ and we have k statistically independent observations, each specified by a standard parametric model $\{p_\theta^{(i)};\ \theta \in \Theta\}$, for i=1, ..., k. The $p_\theta^{(i)}$'s are probability mass functions on a sample space $\mathcal{X}$. Each $p_\theta^{(i)}$ describes a different statistical experiment, but all are governed by the same parameter Θ. For the remote sensing problem, $p_\theta^{(i)}$ could be the model for multispectral scanner (MSS) data, and $p_\theta^{(2)}$ could be the model for the elevation data, etc. Each observation $x^{(i)}$, i=1,...,k gives rise to a belief function, $\mathrm{Bel}^{(i)}_{x(i)}(\theta)$, i=1,...,k, over Θ. $\mathrm{Bel}^{(i)}_{x(i)}(\theta)$, i=1,...,k are constructed depending only on the observation $x^{(i)}$ and the model values $p_{\theta_1}(x^{(i)}), \ldots, p_{\theta_n}(x^{(i)})$ Prior information also gives rise to a belief function, $\mathrm{Bel}_o(\theta)$, over Θ Then the overall belief function is constructed as

$$\mathrm{Bel}\,(\theta) = \mathrm{Bel}_o\,(\theta) \oplus \mathrm{Bel}^{(1)}_{x^{(1)}}\,(\theta) \oplus \mathrm{Bel}^{(2)}_{x^{(2)}}\,(\theta) \oplus \ldots. \oplus \mathrm{Bel}^{(k)}_{x^{(k)}}\,(\theta) \quad , \theta \in \Theta \,. \qquad (2.61)$$

The main conclusions are (Krantz and Miyamoto [25] and Walley [44]) that Dempster's rule is not generally suitable for combining evidence from independent statistical observations (or otherwise, statistically related observations) nor is it suitable for combining prior belief with observational evidence. Stated differently, if the Bel function is interpreted as lower betting rates, then the use of Dempster's rule to combine prior and likelihood functions can lead to a *sure loss* or " Dutch book". That is, Bel cannot *coherently* be interpreted as lower betting rates when Dempster's rule is used to combine priors and likelihood functions.

Finally, it is also interesting to note (Shafer [36]) that in the Bayesian frame of inference, for a given prior probability distribution π_0 over Θ and a given statistical model $\{p_\theta; \theta \in \Theta\}$ over $\mathcal{X}$, one can construct a *unique* distribution p over $\Theta \times \mathcal{X}$, unique in the sense that p is the only distribution that has π_0 as its marginal for θ and p_θ as its conditional given θ. In the D-S theory, there may be *many* belief functions over $\Theta \times \mathcal{X}$ having a given marginal Bel_0 and given the conditional p_θ.

2.5.3 Conditioning rules

An important issue in decision-making and inference is how to change our belief concerning a particular event in light of new evidence. Of course, when the available information is in the form of classical point-valued probabilities Bayes rule provides a natural and sound way of accomplishing this task. In the following section other possibilities are examined.

2.5.3.a Conditional Bel and Pl

Suppose the available information can be represented by a Shaferian belief function, Bel, and plausibility function, Pl, on the frame of discernment Θ. Suppose, further, that somehow one learns that Θ is restricted to B, $B \subset \Theta$. Then Shafer [35] suggests the following:

1) Represent this new information as a new belief function

$$\text{Bel}(A) = \begin{cases} 1 & \text{if } B \subset A \\ 0 & \text{otherwise.} \end{cases} \tag{2.62}$$

2) Combine this belief function with the belief function available prior to the new information by the Dempster's rule of combination to get

$$\mathrm{Bel}\,(A \mid B) = \frac{\mathrm{Bel}\,(A \cup B^c) - \mathrm{Bel}\,(B^c)}{1 - \mathrm{Bel}\,(B^c)}\,. \tag{2.63}$$

Since

$$\mathrm{Pl}\,(A) = 1 - \mathrm{Bel}\,(A^c) \tag{2.64}$$

One obtains,

$$\mathrm{Pl}\,(A \mid B) = \frac{\mathrm{Pl}\,(A \cap B)}{\mathrm{Pl}\,(B)}\,. \tag{2.65}$$

2.5.3.b Conditional "sup" and "inf"

Referring to section 2.3, suppose imprecision about the available information is represented by a family of additive probability distributions $\mathcal{P}$; and

$$P^*(A) = \sup_{P \in \mathcal{P}} P(A) \tag{2.66}$$

and

$$P_*(A) = \inf_{P \in \mathcal{P}} P(A) \tag{2.67}$$

Suppose the new information implies that Θ is restricted to B, $B \subset \Theta$. Then one natural way of revising our earlier beliefs (probabilities) is to say

$$P^*(A \mid B) = \operatorname*{Sup}_{P \in \mathcal{P}} \frac{P(A \cap B)}{P(B)} \tag{2.68}$$

and

$$P_*(A \mid B) = \inf_{P \in \mathcal{P}} \frac{P(A \cap B)}{P(B)} \tag{2.69}$$

The following theorem is due to Huber [18] (also see Kyburg [26])

THEOREM 2.3: (REPRESENTATION THEOREM)

Given a belief function, there exists a closed convex set of classical probability function $\mathcal{P}_c$ defined over atoms of Θ such that for every subset A of Θ

$$\mathrm{Bel}(A) = \inf_{P \in \mathcal{P}_c} P(A) \qquad (2.70)$$

And conversely, if $\mathcal{P}_c$ is a closed convex set of classical probability function defined over atoms of Θ, *and* for every $A_1, A_2, \ldots, A_n \subset \Theta$,

$$\inf P(A_1 \cup A_2 \cup \ldots \cup A_n) \geq \sum_{i=1}^{n} \inf P(A_i) - \sum_{i<j} \inf P(A_i \cap A_j) + \ldots . + (-1)^n \inf P(A_1 \cap A_2 \cap \ldots\ldots \cap A_n) . \qquad (2.71)$$

Then there exists a belief function, Bel, such that

$$\mathrm{Bel}(A) = \inf_{P \in \mathcal{P}_c} P(A) \qquad (2.72)$$

Using the above representation theorems, it can be easily shown that

$$\inf_{P \in \mathcal{P}_c} P(A \mid B) \leq \mathrm{Bel}(A \mid B) \leq \mathrm{Pl}(A \mid B) \leq \sup_{P \in \mathcal{P}_c} P(A \mid B) \qquad (2.73)$$

That is, Shafer's rule of conditioning provides a tighter bound on the conditional values.

It is also interesting to note that Bel (A | B) and Pl (A | B) obtained from Shafer's rule of conditioning are still ∞-monotone and ∞-alternating capacities. Shafer's results are questionable, however, since they are directly based on Dempster's rule of combination.

Diaconis and Zabell [12,13] recommend the following rule :

$$P_*(A \mid B) = \frac{P_*(A \cap B)}{P_*(B)} \qquad (2.75)$$

and

$$P^*(A \mid B) = 1 - P_*(A \mid B) . \qquad (2.76)$$

Again, $P_*(\cdot \mid B)$ and $P^*(\cdot \mid B)$ would still be ∞-monotone and ∞-alternating capacities.

2.5.4.c. Proposed conditioning rule

Both Dempster's rule of conditioning (eq.2.65) and Diaconis and Zabell's rule (eq.2.75) are counter-intuitive. For instance, let us consider Dempster's rule. Applying the representation theorem (Theorem 3.2) to the left hand side of the equation one can write

$$P_*(A \mid B) = \inf_{\mathcal{P}} P(A \mid B) = \inf_{\mathcal{P}} \frac{P(A \cap B)}{P(B)} \tag{2.77}$$

Applying Theorem 3.2 to the right hand side of eq. 2.75 one obtains

$$\frac{P_*(A \cap B)}{P_*(B)} = \frac{\inf P(A \cap B)}{\inf_{\mathcal{P}} P(B)} \tag{2.78}$$

But obviously, in general,

$$\inf_{\mathcal{P}} \frac{P(A \cap B)}{P(B)} \neq \frac{\inf P(A \cap B)}{\inf_{\mathcal{P}} P(B)} \tag{2.79}$$

Considering this discrepancy, the following conditioning rule is suggested.

$$P_*(A \mid B) = \frac{P_*(A \cap B)}{P^*(B)} \tag{2.80}$$

and

$$P^*(A \mid B) = 1 - P_*(A^c \mid B) \tag{2.81}$$

Notice that our definition (eq. 2.80) differs from, for instance, eq. 2.75 in that lower conditional probabilities are computed as ratio of lower joint probabilities and *upper* marginals; that is, the normalizing factor in the denominator is $P^*(B)$ instead of $P_*(B)$.

It may be shown (proof omitted here) that $P_*(A|B)$ and $P^*(A|B)$ obtained above by our rule of conditioning are also ∞-monotone and ∞-alternating capacities, respectively.

2.6 Problems to be solved

It is important to realize that the representation theorem, Theorem 3.2, states only the existence of a family of probability distributions $\mathcal{P}$. It does not, however, suggest a method of constructing $\mathcal{P}$, nor does it imply the *uniqueness of* $\mathcal{P}$.

Our attempt here is in two directions: 1) Try to remedy the problems, mentioned earlier, with the Dempster's rule; that is, the main effort here is to construct a *Bayes-like* rule for *capacities*. Suggestions for a new rule were made above. Properties of this new rule need further investigation. 2) Try to come up with computationally simple methods of constructing $\mathcal{P}$ so that the powerful tools of Bayesian methods could be applied, even with the imprecise probabilities.

CHAPTER 3

SET-VALUED MEASURES

3.1 Introduction

One of the major criticisms to the Bayesian approach for inference and decision-making is its requirement of precise probability values. It has been argued by many people that prior probabilities are subjective and thus it would be unrealistic to assign crisp and precise values.

Two possible solutions to this problem were distributionally robust approach and the Dempster-Shafer theory. Even though, robust approaches are conceptually easy and appealing, obtaining closed form solutions is usually very difficult, except perhaps for certain type of neighborhoods. Also, the solution is really a " worst-case" type solution.

The belief (and plausibility) functions of the D-S theory being monotone (and alternating) capacities of infinite order, are generalization of "classical" measure; but the theory is mainly constructed around Dempster's rule of combination. In our opinion any theory of statistical inference which is based on Dempster's rule of combination would have serious problems and should be abandoned.

A more natural solution would be to generalize classical measure theory, so that measures instead of taking values in $\mathcal{R}$or $\mathcal{R}^n$, take values in subsets of $\mathcal{R}$or $\mathcal{R}^n$, i.e., $\mathcal{P}(\mathcal{R})$or $\mathcal{P}(\mathcal{R}^n)$.

3.2 Set-valued measures

A set-valued measure was introduced by Artstein [1].(Actually, earlier related work was done by Debru and Schmeidler [6]). A set-valued measure (SVM) is a σ-additive set-function which takes on values in the nonempty subsets of a Euclidean space. Let $(\Omega,\mathcal{A})$ be a measurable space, and $K(\mathcal{R}^n)$ be the nonempty compact subsets of $\mathcal{R}^n$. Then a SVM is defined as,

DEFINITION : A set-valued measure is a set function,

$$\mu : \mathcal{A} \rightarrow K(\mathcal{R}^n) \tag{3.1}$$

with the following properties:

(1) $\mu(\emptyset) = 0$

(2) $\mu(\cup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} \mu(A_j)$, for every disjoint family $\{A_j\}_j$, $A_j \in \mathcal{A}$.

where the summation above, is a series of compact subsets of $\mathcal{R}^n$. The sum $\sum_{j=1}^{\infty} \mu(A_j)$ of the subsets $\mu(A_j)$, consists of all the vectors $a=\sum_{j=1}^{\infty} a_j$ where the series is absolutely convergent, and $a_j \in \mu(A_j)$ for $j=1,2,\ldots$.

The interval-valued probability measure (IVPM) Φ (see Negoiwta and Ralescu [28]) is a special type of SVM defined as

$$\Phi : \mathcal{A} \rightarrow K([0,1]) \tag{3.2}$$

and satisfying the properties

(1) $1 \in \Phi(\Omega)$;

(2) $\Phi(\cup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} \Phi(A_j)$.

where $\cup A_j$ is disjoint collection of events in $\mathcal{A}$ and the summation is as defined earlier.

Example 1: Suppose $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and the (objectively, or subjectively) following values are obtained for

$$\Phi(\{\omega_1\}) = [0.6, 0.7]$$
$$\Phi(\{\omega_2\}) = [0.1, 0.15]$$

Then one necessarily gets $\Phi(\{\omega_3\}) = [a, 0.15]$, where $a \leq .15$. Note that,

$$\Phi(\Omega) = \Phi(\omega_1 \cup \omega_2 \cup \omega_3) = \sum_{i=1}^{3} \Phi(\omega_i)$$
$$= [.7+a, 1]$$

and for a =.15, $\Phi(\Omega) = [0.85,1]$.

Remark : Note that $\Phi(\Omega) = [0.85,1] \neq \{1\}(\text{or}[1,1])$. This is also counter-intuitive; we will return to this point shortly.

Negoiwta and Ralescu [28] have shown the following results.

Result 1 : The *conditional probability*, given $M \in \mathcal{A}$ of an event $A \in \mathcal{A}$ is given by

$$\Phi(A \mid M) = \frac{1}{\sup \Phi(M)} \Phi(A \cap M)\,. \tag{3.3}$$

Result 2: $\Phi(M|A)$ and $\Phi(A|M)$ are related by,

$$\Phi(M \mid A) = \frac{\sup \Phi(M)}{\sup \Phi(A)} \Phi(A \mid M) \tag{3.4}$$

and most importantly, the Bayes formula for the interval-valued probability measures (IVPM) is given by

THEOREM :
Let $A_1, A_2, \ldots, A_n$ form a partition of the sample space Ω, and let $B \in \mathcal{A}$ be an event. Then

$$\Phi(A_i \mid B) = \frac{\sup \Phi(A_i)}{\sum_{j=1}^{n} \sup \Phi(A_j) \sup \Phi(B \mid A_j)} \Phi(B \mid A_i) \quad , \; i=1,2, \ldots.,n\,. \tag{3.5}$$

Returning to the problem of statistical inference and decision-making, let $\mathcal{X}$ be the sample space of outcomes (or data), and Θ be a finite parameter space, i.e. $\Theta=\{\theta_1,\ldots,\theta_n\}$. Let π_o be an interval-valued prior probability measure on Θ and $\{\Phi_\theta(x);\ \theta \in \Theta\}$ be a family of conditional interval-valued probability measures on $\mathcal{X}$. Then the Bayes theorem above can be restated as

$$\pi(\theta_i \mid x) = \frac{\sup \pi_o(\theta_j)}{\sum_{j=1}^{n} \sup \pi_o(\theta_j) \sup \Phi(x \mid \theta_j)} \Phi(x \mid \theta_i)\,. \tag{3.6}$$

Then upon observing x, the inference or decision-making could be based on the interval-valued *posterior* probability measure $\pi(\theta_i|x)$.

Example 2 : Suppose $\mathcal{X} = \{ x_1, x_2 \}$, and $\Theta = \{ \theta_1, \theta_2 \}$ and the values of π_o and $\{\Phi_\theta, \theta \in \Theta\}$ are as

$$\pi_o(\theta_1) = [.5, .6]$$

$$\pi_o(\theta_2) = [.2, .4]$$

and

$$\begin{cases} \Phi(x_1 | \theta_1) = [.1, .3] \\ \Phi(x_2 | \theta_1) = [.6, .7] \end{cases} \quad \text{and} \quad \begin{cases} \Phi(x_1 | \theta_2) = [.8, .9] \\ \Phi(x_2 | \theta_2) = [0, .1] \end{cases}$$

Then,

$$\begin{aligned} \pi(\theta_1 | x_1) &= \frac{.6}{(.6)(.3) + (.4)(.9)} [.1, .3] \\ &= [.11, .33] \\ \pi(\theta_1 | x_2) &= \frac{.6}{(.6)(.7) + (.4)(.1)} [.6, .7] \\ &= [.78, .92] \\ \pi(\theta_2 | x_1) &= \frac{.4}{(.6)(.3) + (.4)(.9)} [.8, .9] \\ &= [.59, .67] \\ \pi(\theta_2 | x_2) &= \frac{.4}{(.6)(.7) + (.4)(.1)} [0, .1] \\ &= [0, .08]. \end{aligned}$$

The following definitions and theorem are due to Artstein [1] and Puri and Ralescu [32] and will be used in the sequel.

DEFINITION:

An *atom* of the interval-valued probability measure π is an event $A \in \mathcal{A}$ with $\pi(A) \neq \{0\}$ and such that $A_1 \subset A$ implies $\pi(A_1) = \{0\}$ or $\pi\{A \backslash A_1) = \{0\}$. An interval-valued probability with no atoms is called *nonatomic.*

DEFINITION:

A selection ρ of an interval-valued probability measure π is a vector-valued measure $\rho: \mathcal{A} \rightarrow \mathcal{R}^n$, such that $\rho(A) \in \pi(A)$ for every $A \in \mathcal{A}$.

THEOREM:

(i) If π is a bounded, nonatomic set-valued measure, then $\pi(A)$ is convex for every $A \in \mathcal{A}$.

(ii) If π is a bounded set-valued measure, then for every $A \in \mathcal{A}$ and $s \in \pi(A)$, there exists a selection ρ of π such that $\rho(A) = s$.

Note that clearly interval-valued probability measures are bounded.

Remark: For nonatomic interval-valued probability measures, let us denote $p_1(A)=\inf \pi(A)$ and $p_2(A)=\sup \pi(A)$.

COLLORARY:

For a nonatomic interval-valued probability measure, p_2 is a regular probability measure.

proof: This follows from the convexity of π and the requirement of $1 \in \pi(\Omega)$.

A point mentioned earlier and delayed for here is that the above definition of interval-valued probability measure requires $1 \in \pi(\Omega)$, instead of $\pi(\Omega) = \{1\}$; $\pi(\Omega)=[a,1]$ where $a \leq 1$. This seems counter-intuitive because one expects that Ω should happen *almost surely*.

There are perhaps two ways this point may be addressed:

1) Allow the possibility of $\pi(\Omega)=[a,1]$ $a \leq 1$, and interpret the quantity $(1-a)$ as the "degree of uncertainty" about the space of outcomes, Ω.
2) Add the extra requirement that $\pi(\Omega)=1$. But from this requirement, plus the requirement of additivity, under Minkowski set additions, it immediately follows that one may come up with the interval-valued probability of an event A, such that, $\pi(A)=[p,q]$ and $p > q$; i.e., the set of values π takes on may be possibly not an *ordered set.*

If one should insist that π take on values from an ordered set, plus the requirement $\pi(\Omega)=\{1\}$ and the additivity property, then one should replace Minkowski addition with a different type of set addition operation.

Since the main subjects under consideration are inference and decision-making, these issues are addressed next.

CHAPTER 4

INFERENCE AND DECISION-MAKING WITH IMPRECISE POSTERIOR PROBABILITIES

4.1 Introduction

Regardless of the method used to model imprecise prior probabilities and the conditional probabilities, and how they are combined to obtain posterior probabilities, the next issue is how does one proceed with these imprecise posteriors to make inferences and decisions.

In statistical inference the goal is not to make an immediate decision, but instead to provide a "summary" of the statistical evidence which a wide variety of future "users" of this evidence can easily incorporate into their own decision-making process. Posterior probabilities carry the required information. So as far as the statistical inference is concerned, once the posterior probabilities are obtained the task is completed.

In a decision-making process, however, given an observation, prior information and the models (or conditional densities), rationality dictates that an action a_i, from the set of possible actions, should be chosen that has minimum expected loss (risk).

To be more specific, let us assume a countable parameter set Θ, an action set $\mathcal{A}=\{a_1, a_2,\ldots,a_m\}$, an observation set $\mathcal{X}$, and a loss function

$$L : \mathcal{A}\times\Theta \rightarrow \mathcal{R} \tag{4.1}$$

such that $L(a_i,\theta_j)$ is the loss incurred when action a_i is selected and the state of nature (parameter) is a_j ; and the set $\mathcal{D} = \{\delta_1,\delta_2,\ldots\}$ of nonrandomized decision functions

$$\delta : \mathcal{X} \rightarrow \mathcal{A} . \tag{4.2}$$

Note that in many applications (e.g., estimation problems) $\mathcal{A} = \Theta$. Furthermore, let us represent the "posterior" upper and lower "probabilities" obtained from combination of imprecise priors and imprecise model by $\left\{P^*_x(\theta_i) \text{ and } P_{*x}(\theta_i);\ \theta_i\in\Theta,\ x\in\mathcal{X}\right\}$. We put "posterior" and "probabilities" in quotation marks, because these upper and lower quantities may not be posteriors in the Bayesian sense, and most likely would not be probabilities in the classical probability sense; at best, they may be ∞-alternating and ∞-monotone capacities. The question is:

Given $\left\{P^*_x(\theta_i) \text{ and } P_{*x}(\theta_i);\ \theta_i \in \Theta,\ x \in \mathcal{X}\right\}$ how does one compute expected losses?

4.2 How should upper and lower expectations be defined?

Without loss of generality, assume the loss function is a positive function. Then a natural way to define upper and lower expected loss is to define them (analogous to classical probability) as

$$E^* L(a_i,\theta) = \sum_{\{\,k : (\exists\theta)\ \&\ k = L(a_i,\theta)\,\}} k \,.\, P^*_x \{\, \theta' : L(a_i,\theta') = k \,\} \tag{4.3}$$

and

$$E_* L(a_i,\theta) = \sum_{\{k : (\exists\theta)\ \&\ k = L(a_i,\theta)\,\}} k \,.\, P_{*_x} \{\, \theta' : L(a_i,\theta') = k \,\} \tag{4.4}$$

Note that E* [and E*] would be 2 (or higher order) alternating [and monotone] capacities if P* [and P*] are 2 (or higher order) capacities.

Wolfeson and Fine [47], following Dempster, define the upper and lower expectation as

$$E^* L(a_i,\theta) \triangleq \bar{L}(a_i) = \sum_{\{k : (\exists\theta)\ \&\ k = L(a_i,\theta)\,\}} k \,.\left[\, P_{*_x}(\{\, \theta' : L(a_i,\theta') \le k \,\}) - P_{*_x}(\{\, \theta' : L(a_i,\theta') < k \,\}) \right] \tag{4.5}$$

and

$$E_* L(a_i,\theta) \triangleq \underline{L}(a_i) = \sum_{\{k : (\exists\theta)\ \&\ k = L(a_i,\theta)\,\}} k \,.\left[\, P^*_x(\{\, \theta' : L(a_i,\theta') \le k \,\}) - P^*_x(\{\, \theta' : L(a_i,\theta') < k \,\}) \right] \tag{4.6}$$

When P* [and P*] are 2-alternating [and monotone] capacities, E* [and E*] have, among others, the following properties:

1) $(\forall Z)\ E^* Z \geq E_* Z$

2) $E^* (-Z) = - E_* (-Z)$; i.e E^* and E_* are conjugates.

Also if one obtains P* and P* from

$$P^*(A) = \sup_{P \in \mathcal{P}} P(A) \qquad \forall A \in \mathcal{A} \tag{4.7}$$

$$P_*(A) = \inf_{P \in \mathcal{P}} P(A) \qquad \forall A \in \mathcal{A} \tag{4.8}$$

then

$$E^*(Z) = \sup_{p \in \mathcal{P}} E_p(Z) \tag{4.9}$$

$$E_*(Z) = \inf_{p \in \mathcal{P}} E_p(Z) \tag{4.10}$$

Note that the above upper probabilities are used to compute the lower expectations and vice-versa. Note also that upper and lower expectations given by

$$E^* L(a_i,\theta) = \sum_{\{k :(\exists\theta)\ \&\ k= L(a_i,\theta)\ \}} k \,.\, P_{*_x}(\{\theta' : L(a_i,\theta') = k \}) \tag{4.11}$$

$$E_* L(a_i,\theta) = \sum_{\{k :(\exists\theta)\ \&\ k= L(a_i,\theta)\ \}} k \,.\, P_x^* (\{\theta' : L(a_i,\theta') = k \}) \tag{4.12}$$

are different than the ones given in (4.3) and (4.4). Furthermore, since in general

$$P^*(\{\theta : L(a_i,\theta) = k \}) \neq \Big[P^*(\{\theta : L(a_i,\theta) \leq k \}) - P^*(\{\theta : L(a_i,\theta) < k\}) \Big] \tag{4.13}$$

$$P_*(\{\theta : L(a_i,\theta) = k \}) \neq \Big[P_*(\{\theta : L(a_i,\theta) \leq k \}) - P_*(\{\theta : L(a_i,\theta) < k\}) \Big] \tag{4.14}$$

using the right hand side of (4.13) and (4.14) in (4.3) and (4.4) would result in yet different values. Which one of the upper and lower pair of values is correct ? One may have to experimentally justify one pair over the other. Thus given an observation, regardless of which method is used to get the expected values, one obtain a pair of upper and lower expected losses. Then decisions are based on the values of these pairs.

With the usual point-value probabilities, expected losses are also point-valued; and we choose an action that has minimum expected loss (risk). For upper and lower expected losses, however, the problem is a little more complicated.

When the upper and lower expected loss (U&L EL) intervals are non-intersecting, the choice of an action is easy. That is, we order acts by dominance: $a_1 > a_2$ (read a_1 is *preferred* to a_2) if and only if

$$\underline{L}(a_1) > \bar{L}(a_2) \qquad (4.15)$$

And for more than two actions, we choose action a_i* such that

$$a_i^* = \arg(\max_j \underline{L}(a_j)) \qquad (4.16)$$

When the (U&L EL) intervals overlap, however, we face the problem of indecisiveness. When $\underline{L}(a_j) > \underline{L}(a_i)$ but $\bar{L}(a_j) < \bar{L}(a_i)$ (i.e., $[\underline{L}(a_j), \bar{L}(a_j)] \subset [\underline{L}(a_i), \bar{L}(a_i)]_)$, that is intervals are nested, and it is not clear which action should be preferred and why.

What can be done, however, is to eliminate from the set of possible actions, those that are *not* preferable. That is, suppose for a_k, $k \neq i$, $k \neq j$, $k=1,2,\ldots,m$,

$$\bar{L}(a_k) < \underline{L}(a_i)$$

and

$$\bar{L}(a_k) < \underline{L}(a_j).$$

Then eliminate a_k, $k \neq i$, $k \neq j$, $k=1,2,\ldots,m$ from further considerations. And try to resolve the remaining indecision between a_i and a_j. Note also that one may face indecisiveness between a_i and a_j when,

$$L\,(a_j) > L\,(a_i)$$

and

$$\bar{L}\,(a_j) > \bar{L}\,(a_i).$$

There are two possibilities at this point: 1) Claim indecisiveness and require more information (e.g., in the form of more sample data for the frequentist approach), 2) Use some *ad hoc* but "reasonable" approach to resolve the problem. Let us show the above situation graphically (see fig. 4.1).

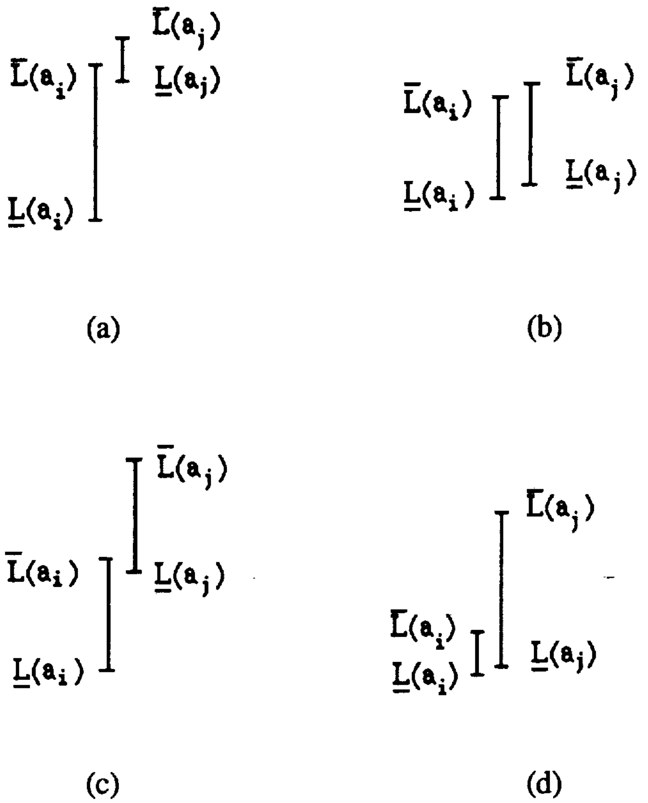

Fig. 4.1 - Four possibilities for actions a_i and a_j with overlapping expected utilities : a) $L(a_j)$much larger than $L(a_i)$ but $\bar{L}(a_j)$ slightly larger than $\bar{L}(a_i)$. b) , c) , d) etc.

In fig. 4.1 above the following is recommended:

For a) $a_j > a_i$ That is a_j is preferred over a_i.

For b) a_j and a_j are about equally preferable; this situation can happen in point-valued expected loss problems too when the expected loss of two actions are equal. We say that we are ***indifferent*** about a_j and a_j ; and use a "tie-breaking" rule to decide.

For c) $a_j > a_{i.}$

For d) Again $a_j > a_{i.}$

CHAPTER 5

5.1 Summary

This work examined the following three issues. First, how best to describe the imprecise knowledge about prior probabilities and conditional densities. Second, how best to combine these imprecise values to get the so called *posterior* probabilities. And finally how to make decisions with imprecise posteriors.

Various methods in the literature such as distributionally robust approaches, Dempster-Shafer theory and set-valued measures were examined. It was noted that even though distributionally robust approaches offer intuitively simple ways of expressing imprecision in the available knowledge, in general obtaining closed form solutions for the minimax decision rules except for some special families of distributions, namely classes of distributions majorized with Choquet capacities, are very difficult. Also robust methods really treat only the problem of designing against the *worst case* situations.

In examining the Dempster-Shafer (D-S) theory, it was noted that even though D-S theory provides a reasonable method for modeling imprecision, there are at least two major problems with the theory: 1) The theory is mainly built around the Dempster's rule of combination of evidences; this rule, however, has been under major criticisms. Recalling that Dempster-Shafer's upper and lower probabilities (or in D-S language, the plausibility and belief) are ∞- alternating and ∞-monotone capacities, respectively, then the main thrust here should be an attempt to find a Bayes-like rule for capacities. 2) The computational complexity of the Dempster's rule is shown to be #P-complete [30]. That is, even given as input a set of tables representing basic probability assignments m_1, m_2, ..., m_n over a frame of discernment Θ, and a set $A \subseteq \Theta$, the problem of computing the basic probability value $(m_1 \oplus m_2 \oplus \dots \oplus m_n)(A)$ is #P-complete.

Interval-valued probabilities (or set-valued measures in general) start from the very beginning by assigning intervals (or sets) to each event. That is, if one is not able to assign single values to the probabilities of events, one will assign intervals (or sets) of values for the probabilities. The real-valuedness axiom of conventional probability theory is relaxed. Then, in an attempt to preserve the (countable) additivity axiom, the additivity is defined in terms of set additions. The main problem here, at least with the current definition of set additions, is that one cannot simultaneously enforce the requirements that: 1) measure of null even has to be zero; 2) keep the additivity axiom; and 3) have the measure of the *sure event* equal to one. Therefore, the third requirement is relaxed. This is, however, quite counter-intuitive since then one could define a new event and assign the remaining probability mass to this event.

Finally, we looked at the issue of decision making with imprecise posterior probabilities. This rises from the fact that if one starts with imprecise models and/or imprecise priors one is bound to arrive at imprecise posteriors. The specific form of the set of the posteriors at this point is irrelevant. Even though some specific situations were considered, the problem basically still remains as an open problem. This is because the conventional decision theory (based on the utility theory) assumes point-valued probabilities. Preferences on the set of actions or decisions are ordered using their expected utilities. It is this ordering property that is lost when we consider sets of imprecise probabilities.

References

[1] Z. Artstein, " Set-valued measures," *Trans. Amer. Math. Soc.* **165**, 103-125 (1972).

[2] R.J. Beran, " Upper and lower risks and minimax procedures," In Proc. Sixth Berkeley Symp. Math. Statist. Probab., Univ. of California Press.

[3] J.O. Berger, " Statistical decision theory and Bayesian analysis," Springer-Verlag, Second edition (1985).

[4] G. Boole, " An investigation of the laws of thought," (1854); Reprinted by Dover (1958).

[5] G. Choquet, " Theory of capacities," *Ann. Inst. Fourier* **5**, 131-295 (1953).

[6] G. Debreu and D. Schmeidler, " The Raydon-Nikodym derivatives of a correspondence," Proc. Sixth Berkeley Symp. Math. Statist. Probab., 41-56. Univ. of California Press.

[7] B. De Finetti, " Probability, induction and statistics," Wiley (1972).

[8] A. Dempster, " A generalization of Bayesian inference (with discussion)," *J. Royal Statist. Soc.* **B** 30, 205-245 (1968).

[9] A. Dempster, " New methods for reasoning towards posterior distributions based on sample data," *Ann. Math. Statist.* **37**, 355-374 (1966).

[10] A. Dempster, " Upper and lower probabilities induced by a multivalued mapping," *Ann. of Math. Statist.,* **38**, 325-329 (1967).

[11] L. DeRobertis and J. A. Hartigan, "Bayesian inference using intervals of measures," *Ann. Statist.* vol. **9**, No. 2, 235-244 (1981).

[12] P. Diaconis and S. Zabell," Updating subjective probabilities," *J. Statist. Assoc.* **77**, 822-830 (1980).

[13] P. Diaconis and S. Zabell," Some Alternatives to Bayes' rule," TR. No. 339 Stanford University (1983).

[14] P.C. Fishburn, " Analysis of decisions with incomplete knowledge of probabilities," *Op. Res.* **13** 217-237 (1965).

[15] J.P. Florens et. al. (eds.), " Specifying statistical models," Lecture Notes in Statistics # **16**, Springer-Verlag (1981).

[16] D.C. Heath and W.D. Sudderth, " On finitely additive priors, coherence, and extended admissibility," *Ann. Statist.* **43** 2072-2077 (1978).

[17] P. Huber and V. Strassen, " Minimax tests and the Neyman-Pearson lemma for capacities," *Ann. Statist.,* **1**, 251-263 (1973).

[18] P. Huber, " The use of Choquet capacities in statistics," *Bull. of the Internat. Statist. Inst. Vol* . **XLV**, Book 4, 181-188 (1973).

[19] E.T. Jaynes, " Prior probabilities," *IEEE Trans. Syst. Sci. Cybern.*, vol. **SSC-4**, 227-241 (1968).

[20] S.A. Kassam, " Robust hypothesis testing for bounded classes of probability densities," *IEEE Trans. Inform. Theory*, Vol. **IT-27**, No. 2, 242-247 (1981).

[21] J.B. Kadane and D.T. Chung, "Stable decision problems," *Ann. Statist.* **6** 1059-1110 (1978).

[22] Y.S. Kharin, " Stability of decision rules in pattern recognition problems," *Automatika and Remote Control,* II 115-123 (1982).

[23] G.J. Klir and T.A. Folger, "Fuzzy sets, uncertainty, and information," Prentice-Hall (1988).

[24] B.O. Koopman, " The axioms and algebra of intuitive probability'" *Ann. Math.* **41** 269-278 (1940).

[25] D.H. Krantz and J. Miyamoto, " Priors and likelihood ratios as evidence," *J. Amer. Statist. Assoc.* **78**, 418-423 (1983).

[26] H.E. Kyburg Jr., " Bayesian and non-Bayesian evidential updating," *Artificial Intelligence* 271-293 (1987).

[27] D.A. Lane and W. D. Sudderth, " Coherence and continuous inference," *Ann. Statist.* Vol. **11**, No. 1, 114-120 (1983).

[28] C.V. Negoiwta and D. Ralescu," Simulation, knowledge-based computing, and fuzzy statistics," Van Nostrand Reinhold (1987).

[29] H.T. Nguyen, " On random sets and belief functions," *Ann. Math. Anal. Appl.* **65**, 531-542 (1978).

[30] P. Orponen," Dempster's rule of combination is #p-complete," *Artificial Intelligence* 245-253 (1990).

[31] J.M. Potter and B.D. Anderson, " Partial prior information and decisionmaking, " *IEEE trans. Syst. Man Cybern.*, Vol. **SMC**-10, No.3, 125-133 (1980).

[32] M.L. Puri and Dan A. Ralescu, " Strong law of large numbers with respect to a set-valued probability measure," *Ann. Prob.*, Vol. **11**, No.4, 1051-1054 (1983).

[33] E. Regazzini, " De Finetti's coherence and statistical inference," *Ann. Statist.* **15** no. 2 845-864 (1978).

[34] W.J. Rey, " Robust statistical methods," Lecture notes in math. 690 (1980).

[35] G. Shafer," A mathematical theory of evidence," Princeton Univ. Press (1976).

[36] G. Shafer," Belief functions and parametric models (with discussion)," *J. Roy. Statist. Soc. Ser.* B **44**, 322-352 (1982).

[37] G. Shafer," Allocation of probability : A theory of partial belief," Princeton doctoral dissertation (1973).

[38] G. Shafer, " A theory of statistical evidence (with discussion)," In *Foundation of Probability Theory, Statistical Inference, and Statistical Theories of Science* (w.L. Harper and C.A. Hooker, eds.) **2**, 365-436 Riedel, Dordrecht.

[39] J.E. Shore and R.W. Johnson, " Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*, Vol. **IT**-26, 26-37 (1980).

[40] J.E. Shore and R.W. Johnson, " Properties of cross-entropy minimization," *IEEE Trans. Inform Theory*, vol. **IT**-27, 472-482 (1981).

[41] J.E. Shore and R.M. Gray, "Minimum cross-entropy pattern classification and cluster analysis," *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. **PAMI**-4, No.1, 11-17 (1981).

[42] C.A.B. Smith, " Consistency in statistical inference and decision (with discussion)," *J. Roy. Statist. Soc. Ser.* **B**, 23, 1-25 (1961).

[43] P. Suppes and M. Zanotti, " On using random relations to generate upper and lower probabilities, " *Synthese* **36** 427-440 (1977).

[44] P. Walley, " Belief function representations of statistical evidence," *Ann. Statist.* Vol. **15**, No. 4, 1439-1465 (1987).

[45] P. Walley and T.L. Fine, " Toward a frequentist theory of upper and lower probability," *Ann. Statist.* **10** no.3 741-761 (1983).

[46] P.M. Williams, " Indeterminate probabilities," In Formal methods in the methodology of empirical sciences, M. Przelecki, K. Szaniawski, and R. Wojeiki (eds.) Reidel (1976).

[47] M. Wolfenson and T.L. Fine, " Bayes-like decision making with upper and lower probabilities," *J. Amer. Statsit. Assoc.* **77**, 80-88 (1982).

[48] L.A. Zadeh," On the validity of Dempster's rule of combination," memorandum No. UCB/ERL M79/24, Univ. of Calif., Berkeley (1979).

[49] L.A Zadeh," Review of : 'A mathematical theory of evidence' by G. Shafer," *Artificial Intelligence,* 81-83 (1984).

[50] L.A. Zadeh," A simple view of The Dempster-Shafer theory of evidence and its implication for the rule of combination," *Artificial Intelligence,* 85-90 (1986).